INTEGRATION the VLSI journal 58 (2017) 421–429

Contents lists available at ScienceDirect

INTEGRATION, the VLSI journal

journal homepage: www.elsevier.com/locate/vlsi

Thermal aware design and comparative analysis of a high performance 64-bit adder in FD-SOI and bulk CMOS technologies

Can Baltacı*, Yusuf Leblebici

LSM, STI, EPFL, 1015 Lausanne, Switzerland

ARTICLE INFO

Keywords:
Self-heating
Thermal modelling
Thermal simulation
Bulk vs FDSOI
64-bit adder in dynamic logic

ABSTRACT

Thermal behaviours of high-performance digital circuits in bulk CMOS and FDSOI technologies are compared on a 64-bit Kogge-Stone adder designed in 40 nm node. Temperature profiles of the adder in bulk and FDSOI are extracted with thermal simulations and hotspot locations are studied. The influence of local power density on peak temperature is examined. It is shown that high power density devices have significant influence on peak temperature in FDSOI. It is found that some group of devices that perform the same function are the most prominent heat generators. A modification on the design of these devices is proposed which decreases the hotspot temperatures significantly.

1. Introduction

The demand for increasing the performance of high speed digital circuits brings the need for smaller and faster implementations [1,2]. However, as the clock frequencies increase owing to smaller technology nodes, the circuits consume power in higher densities. Consequently, the temperature levels are elevated and the thermal issues become the bottleneck of the circuits by altering the performance and decreasing the lifetime. On the performance side, the temperature induced reduced mobility decreases the device current and the maximum speed of operation [3]. Moreover, the threshold voltage decreases with the temperature and this results in higher leakage current and power consumption [4,5]. Higher power consumption brings higher temperature and this might result in thermal runaway where the die fails due to the uncontrolled increase in the temperature. Although thermal runaway does not happen, the chip might settle down to a higher temperature, which would degrade the performance as well as the reliability of the chip [6]. Electromigration phenomena is another reliability problem related to temperature where the metal interconnects are broken due to diffusion or flow of atoms under very high current densities at high temperatures [7,8]. All of the mentioned problems show that having a reliable and high performance chip is not possible without considering the thermal behaviour of the design. This brings another aspect into the design space, which is the self-heating.

Self-heating became a more critical problem especially after the introduction of the modern MOSFET device geometries like FinFET and Fully Depleted Silicon on Insulator (FDSOI) [9]. Previously, it was reported that the peak temperature of the FDSOI devices is located close to the drain end of the device [1,10,11] and the peak temperature value in FDSOI FETs is found to be much higher than the one in the conventional bulk MOSFETs [12,13]. The higher peak temperature of the FDSOI structure is mainly due to the thermal behaviour of its constitutive materials. The thermal conductivity of the SiO_2 isolation layer is two orders of magnitude lower than the thermal conductivity of the bulk Si. Moreover, the thermal conductivity of Si thin film, where the devices are generating heat, is one order of magnitude less than the thermal conductivity of bulk Si [1]. Additionally, the boundary between Si and SiO_2 creates a finite interface thermal resistance, [14,15] which is equal to the thermal resistance of a SiO_2 layer with a thickness of 20 nm [16]. Due to the mentioned facts, the dissipated power in FDSOI devices does not find a high conductance diffusion path. Consequently, the generated heat turns into temperature in nanometer scale local spots which are comparable to the dimensions of transistors in FDSOI. However, not all the devices settle down to very high temperature values in an implementation. The devices which consume the highest amount of power per unit area are the hottest ones especially in FDSOI. As a result of this, a design which contains devices with large differences in their power densities create very prominent temperature hotspots and large temperature gradients. By performing a detailed power density analysis, the critical ones can be eliminated from the others; and by performing some modifications on their design, the peak temperature and the high temperature gradients can be reduced. Recently, during the implementation of a 5 GHz processor, high switching factor nets were identified during functional simulation to avoid micro hotspots at the individual gate level, caused by device self-heating [17]. As a solution, the maximum output load capacitances of

* Corresponding author.
E-mail addresses: can.baltaci@epfl.ch (C. Baltacı), yusuf.leblebici@epfl.ch (Y. Leblebici).

http://dx.doi.org/10.1016/j.vlsi.2017.03.001

Available online 10 March 2017
0167-9260/ © 2017 Elsevier B.V. All rights reserved.

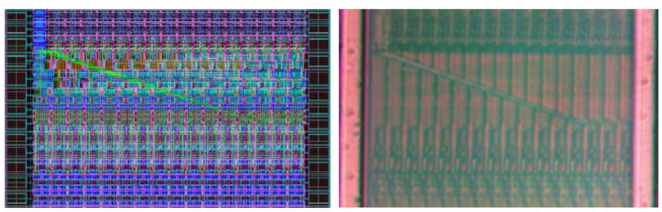
the gates driving these nets are reduced and these gates are placed away from the other gates driving such nets in order to avoid excessive heating and have uniform temperature distribution overall the circuit.

In this work, we intend to emphasize the correlation of the nanometer scale hotspots and the power density of individual devices by observing the temperature profile of high performance circuits implemented in bulk and FDSOI. For this purpose, a 64-bit parallel prefix adder is designed and implemented in a commercially available 40 nm CMOS bulk technology. The power dissipation of each device in the circuit is observed under randomly applied input vectors. The resulting power dissipation output is provided as an input to thermal simulations for observing the temperature profile of the overall block. HotSpot tool [18] is used for modelling the thermal behaviours of bulk and FDSOI geometries. The devices situated on the highest temperature locations are found and examined. It is observed that self-heating in FDSOI is much more prominent when compared to bulk since the local hotspots have sizes comparable to the size of the devices and the generated heat is directly converted to temperature in FDSOI. Consequently, the highest temperature values occur on the devices which have the highest power density. Finally, a solution for decreasing the temperature of the hotspots is proposed. It is shown that the peak temperature of the design in FDSOI can be decreased significantly with a cost of an insignificant increase in the area and parasitic capacitances.

In Section 2, the performance parameters of the implemented 64-bit parallel prefix adder are given and its architecture is explained in detail. In Section 3, bulk and FDSOI thermal simulation results and the temperature profiles of the designed 64-bit adder is provided. In Section 4, the correlations between the devices with the peak temperature and their functions are shown. Finally, in Section 5, the summary of the work and the conclusions are provided.

2. Implemented block

The parallel prefix adder is implemented with Kogge-Stone technique [19] where radix-4 and sparsity-4 options are used [20,21]. The entire 64-bit Kogge-Stone adder block is designed with full custom design approach (Fig. 1). The block is primarily optimized to obtain the lowest possible critical path delay while having the lowest possible power consumption and area. Finally, a delay (clock to sum) of 148 ps is obtained under 900 mV power supply voltage. The block contains 10922 nMOS and pMOS devices and the resulting area of the block is around 2200 μm^2. The average power dissipation of the block is 12 mW and the average power density is 548 W/cm^2 under a clock frequency of 2.5 GHz with 50% duty cycle. This corresponds to 200 ps evaluation time which is 52 ps more than the critical path delay. The detailed block diagram of the implemented 64-bit Kogge-Stone Parallel Prefix Adder can be seen on Fig. 2. The interconnect lines, signal names and the blocks on the critical path are indicated by red colour. The block consists of three main building blocks which are Propagate-Generate Signal Generator, Propagate-Generate Signal Merge and *4-bit* Carry Select Adders (CSA). The detailed explanation of the architecture of these blocks are given in the following sub-sections.



(a) Layout (b) Micrograph

Fig. 1. Full custom layout and die micrograph of the 64-bit adder block (54.8 μm×40 μm) implemented in 40 nm technology. (a) Layout (b) Micrograph.

2.1. Propagate-generate signal generator

In parallel prefix adders, the reduction in the delay time is provided by merging *Propagate* (*P*) and *Generate* (*G*) signals in a parallel fashion to obtain the values of the *Carry Out* signals. For that reason, *Propagate* and *Generate* signals are generated before the merging step according the Boolean expressions given by

$$\overline{P_i} = \overline{A_i + B_i} \tag{1}$$

$$\overline{G_i} = \overline{A_i \bullet B_i} \tag{2}$$

where the subscript i is the bit number of the input values A and B. In this implementation, instead of P and G signals, and signals are generated in the *Propagate-Generate Signal Generator* block, mainly for decreasing the logic depth and increasing the clock frequency. The logic gates are implemented with N-Domino Logic with the clocked footing devices [22].

2.2. Propagate-generate signal merge

The *Propagate-Generate Signal Merge* block is the heart of the overall 64-bit Adder implementation since the evaluation time of this part has an important influence on the speed of the overall block. In this block, P and G signals are merged to get the *Carry Out* information of different stages in the addition. On Fig. 2, the blue coloured circles indicate a logic gate which performs the merging operation of four P and four G signals, where the radix option is set to 4 for further decreasing the critical path delay by decreasing the logic depth [20,23]. The Boolean expression of these functions are shown by (3) and (4) where the subscript $i{:}i{-}3$ indicates that the output signals are the merged P and G signals from the bits i to $i{-}3$.

$$P_{i:i-3} = P_i \bullet P_{i-1} \bullet P_{i-2} \bullet P_{i-3} \tag{3}$$

$$G_{i:i-3} = (G_i + G_{i-1} \bullet P_i) + (G_{i-2} + G_{i-3} \bullet P_{i-2}) \bullet P_i \bullet P_{i-1} \tag{4}$$

The radix-4 option provides the advantage of merging four signals with a single logic gate. However, (3) and (4) shows that the implemented CMOS logic gate will be quite complex and it will contain 4 transistors in series. This fact will in turn decrease the speed of the logic gate especially for the advanced technology nodes where the power supply voltages are equal to or below 1 V. Another possibility to implement a radix-4 PG-Merge gate is to cascade two radix-2 PG-Merge gates. Hence, (3) and (4) can be written as

$$P_{i:i-3} = (P_i \bullet P_{i-1}) \bullet (P_{i-2} \bullet P_{i-3}) = P_{i:i-1} \bullet P_{i-2:i-3} \tag{5}$$

$$G_{i:i-3} = [G_i + G_{i-1} \bullet P_i] + [(G_{i-2} + G_{i-3} \bullet P_{i-2}) \bullet (P_i \bullet P_{i-1})] = G_{i:i-1} + G_{i-2:i-3} \bullet P_{i:i-1} \tag{6}$$

where the cascading of two radix-2 PG-Merge gates can be explicitly seen. The cascading approach has the disadvantage of having a logic depth of two when compared to the approach shown by (3) and (4); however, the series resistance in each gate is quite relaxed. At this point, the delay performance can be questioned since both approaches have their own advantages and disadvantages. To observe the faster solution, the same 64-bit Adder is implemented with the both approaches. It is observed that the cascading approach is unequivocally faster than the single gate approach.

As indicated in Section 2.1, only and signals are available at the inputs of the *Propagate-Generate Signal Merge* block. Consequently, (7) and (8) are used for implementing the radix-4 PG-Merge gates where both the inputs and the outputs are negated.

$$\overline{P_{i:i-3}} = \overline{(\overline{P_i} + \overline{P_{i-1}}) \bullet (\overline{P_{i-2}} + \overline{P_{i-3}})} = \overline{P_{i:i-1} \bullet P_{i-2:i-3}} \tag{7}$$

$$\overline{G_{i:i-3}} = \{\overline{G_i \bullet (G_{i-1} + \overline{P_i})} + [\overline{G_{i-2} \bullet (G_{i-3} + \overline{P_{i-2}})} \bullet \overline{(\overline{P_i} + \overline{P_{i-1}})}]\}' = \overline{G_{i:i-1} + G_{i-2:i-3} \bullet P_{i:i-1}} \tag{8}$$

All radix-4 PG Merge gates in the *Propagate-Generate Signal*

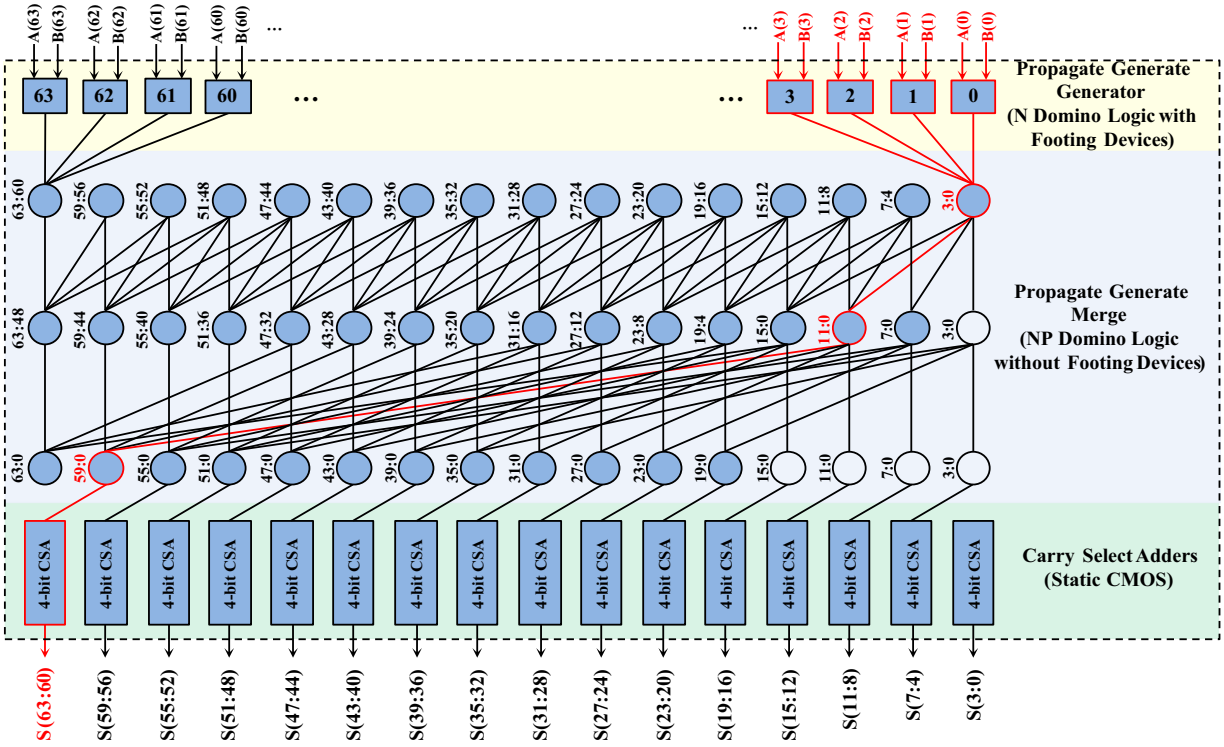


Fig. 2. The architectural block diagram of the implemented 64-bit Kogge-Stone Parallel Prefix Adder with the radix-4 and sparsity-4 options. The input/output nodes, signal names and the hardware blocks on the critical path are indicated with red colour. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

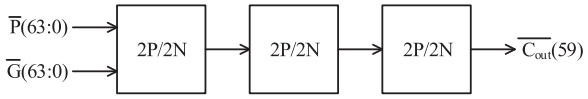


Fig. 3. The number and type of the devices on the critical path of the *Propagate-Generate Signal Merge* block.

Merge block (the blue circles on Fig. 2) are implemented with NP-Domino Logic gates. Since the logic gates in the *Propagate-Generate Signal Generator* blocks are implemented with N-type logic tree, the implementation is proceeded with P-type domino gates to prevent the irreversible discharge problem during evaluation. No clocked footing device is used since there is no possible current path on the evaluation trees of the logic gates. The fact that the clocked footing devices are removed increases the rise and fall times of the gates and decreases the area of the standard cells significantly. The number and type of the transistors on the critical path are summarized on Fig. 3.

2.3. Carry select adders (CSA)

For obtaining the sum value, *Carry Select Adders* are needed since the *Propagate-Generate Signal Merge* block generates only the *Carry Out* signals. Moreover, each carry select adder has to be a 4-bit implementation (Fig. 4) because the *Propagate-Generate Signal Merge* block is implemented with sparsity-4 option where only every forth *Carry Out* signal is generated [21]. The 4-bit carry select adder is implemented with two 4-bit ripple carry adders. In the implementation, *P* and *G* signals are used as the inputs of 1-bit full adders instead of *A* and *B* input bits. This way, it is possible to decrease the number of the series transistors from 3 to 2 in the 1-bit full adder gates. The implementation of the *Carry Select Adder* block is performed by static CMOS since it is not time critical.

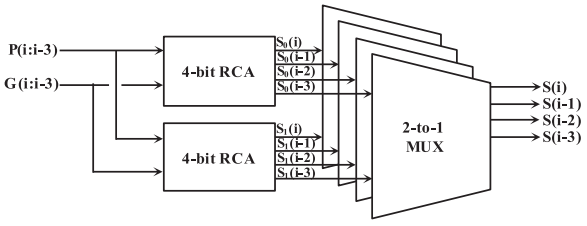


Fig. 4. The block diagram of the 4-bit *Carry Select Adder*. The $S_0(i)$ and $S_1(i)$ signals at the outputs of the two 4-bit ripple carry adders are the *Sum* signals which are generated for the two cases where *Carry In* signal is equal to 0 and 1 respectively.

3. Self-heating analysis

Detailed self-heating analyses are performed on the implemented block by considering the thermal properties of both bulk and FDSOI device geometries. Firstly, electrical simulations are performed where series of randomly generated input vectors are applied at the inputs for extracting the power dissipation waveforms of each device in the block. The instantaneous power dissipation waveform of each device is extracted by

$$P_n(t) = I_{D,n}(t) \times V_{DS,n}(t) \tag{9}$$

and the average power dissipation values are calculated by

$$\overline{P_n} = \frac{1}{t_s}\int_0^{t_s} I_{D,n}(t) \times V_{DS,n}(t)\,dt = \overline{Q_n} \tag{10}$$

where P_n is the instantaneous power dissipation (in Watts), $I_{D,n}$ is the drain current (in Amperes), $V_{DS,n}$ is the voltage drop between the drain and the source terminals (in Volts), and are the average power dissipation and heat generation (in Watts) respectively of device n throughout the simulation time, t_s (in seconds). Finally, the location of each device is extracted from layout for generating the heat generation map of the 64-bit adder and performing thermal simulations.

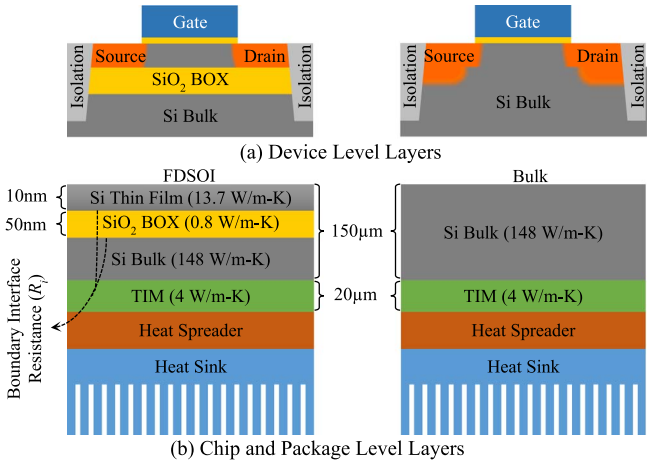


Fig. 5. Thermal models of different layers for FDSOI and bulk (a) Device Level Layers(b) Chip and Package Level Layers.

The electrical simulations are followed by thermal simulations. For the thermal simulations, bulk and FDSOI technologies are separately modelled to make a comparison between each other. For the bulk case, the geometry of a commercially available 40 nm technology is thermally modelled. For the FDSOI case, the technology parameters of a commercially available 28 nm FDSOI design kit are selected. To perform a fair comparison, the parameters of the chosen 28 nm FDSOI technology are scaled for 40 nm. The extracted thermal models for different layers of the chip are illustrated on Fig. 5. The bulk thickness is set to 150 μm for both cases. For the FDSOI case, the Si thin film and the BOX (buried oxide) thicknesses are set to 10 nm and 50 nm respectively. The thermal conductance of the Si and SiO_2 thin film structures are less than their bulk thermal conductance values due to phonon boundary scattering [24]. Therefore, these two layers are independently modelled. The thermal conductance of the SiO_2 film with 50 nm thickness is approximately equal to 0.8 $Wm^{-1} K^{-1}$ [25,26]. The thermal conductance of the Si film with 10 nm thickness is approximately 13.7 $Wm^{-1} K^{-1}$. The calculated thermal conductance value is more than an order of magnitude less than the thermal conductance of bulk Si. The boundary between Si and SiO_2 creates a temperature jump, hence a finite interface thermal conductance [14,15]. The value of interface thermal resistance (R_i) is set to 2×10^{-8} $m^2 KW^{-1}$ [25,26] (Fig. 5). The resulting interface thermal resistance is equivalent to the thermal resistance of a 20 nm thick SiO_2 film. All of the mentioned modifications for the FDSOI case increases the thermal resistance at the top surface significantly. For the FDSOI case, the transistors are located inside the Si Thin Film layer, therefore the extracted heat generators are placed in this layer according to their coordinate information extracted from the layout.

For the bulk Silicon simulations, the heat sources are placed on the top surface of the Si Bulk layer. It can be seen from Fig. 5 that the metal routing layers and the dielectric isolation layer is ignored. Extracting the thermal model of this layer is very difficult since the distribution of the routing layers are quite complex and not homogeneous. However, the average thermal conductance of the metal-dielectric layer is quite low when compared to the bottom heat conduction path (through heat spreader and heat sink). Moreover, the heat conduction of the metal-dielectric layer is even less in deep sub-micron technologies due to the low thermal conductance of the low-κ materials [27]. Therefore, the upper part of the chip is assumed to be thermally insulator.

For providing the calculated power dissipation values as the inputs for the thermal simulations, the block is partitioned into squares (pixels) of 1 μm^2 and the total heat generation of each square is calculated by summing the power dissipation of each device which falls in that square. The resulting heat generation map can be observed on Fig. 6. For observation purposes, a block level heat generation profile is

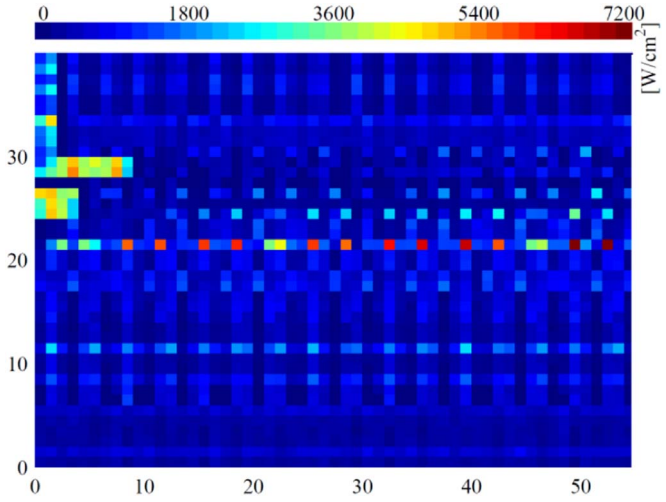


Fig. 6. Heat generation map of the 64-bit adder (x and y in μm). The spatial resolution used for thermal simulation is 1 μm^2. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article).

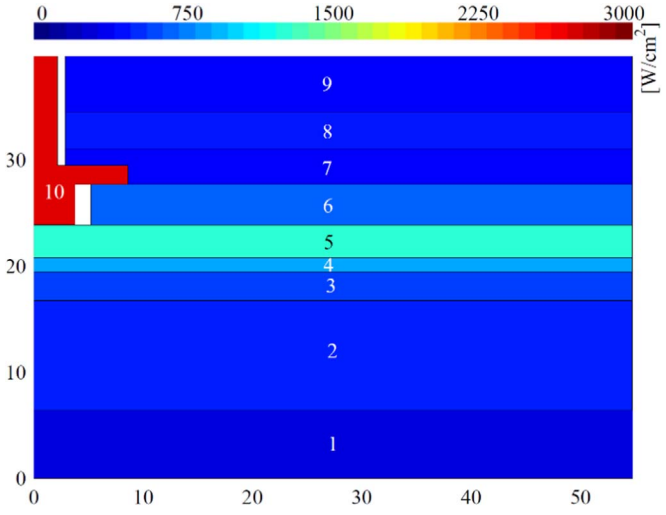


Fig. 7. Heat generation profile of separate blocks (x and y in μm).

Table 1
Functions and heat density values of the blocks on Fig. 7.

Number	Function	Heat density [W/cm^2]
1	Input registers	278.8
2	Carry select adder	469.7
3	PG generators	557.8
4	Inverters	867.4
5	PG Merge 1	1254
6	PG Merge 2	659.0
7	PG Merge 3	365.7
8	Multiplexers	440.2
9	Output registers	369.4
10	Clock Tree	2715

also generated and can be seen on Fig. 7, where the names of the blocks and their individual heat generation values are reported on Table 1. By observing Fig. 7, it can be seen that the block with the highest heat generation density is the *Clock Tree*, which is an expected result when the density of the devices and their activity rates are considered. The effect of the *Clock Tree* can also be seen by observing the pixels on the upper left corner of Fig. 6.

However, the pixels with the highest heat generation density on Fig. 6 do not belong to the *Clock Tree*. The pixels with highest heat

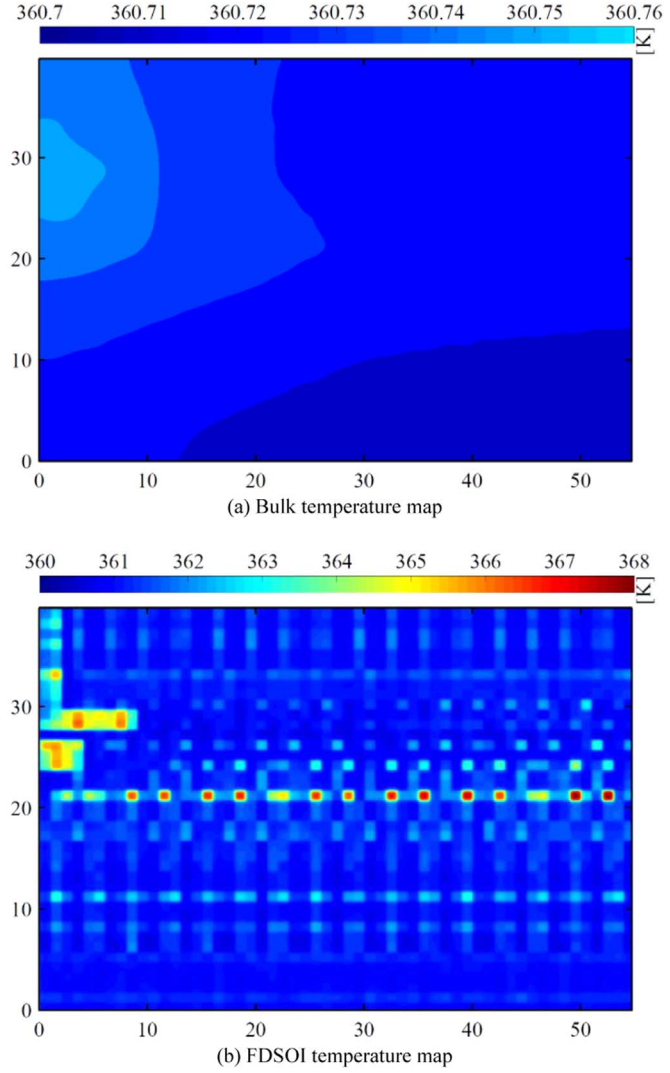


Fig. 8. Temperature profiles of the 64-bit adder (x and y dimensions are in μm). The spatial resolution used for thermal simulation is 1 μm^2. (a) Bulk temperature map (b) FDSOI temperature map.

generation density are found between y=21 μm and y=22 μm and located in the *Propagate-Generate Merge* block. Nevertheless, the average heat generation rate of the *Propagate-Generate Merge* block is still less than the *Clock Tree* due to the fact that not all of the devices in this block dissipate heat as much as the ones inside the dark red pixels which can be seen on Fig. 6.

The heat generation map of Fig. 6 rather than the block level one of Fig. 7 is applied as the input of the two thermal simulations which are performed by using bulk and FDSOI thermal models separately. The temperature maps which are obtained from the thermal simulations for the two cases are shown on Fig. 8. It is clear that the temperature profiles of bulk and FDSOI structures differ significantly. For the bulk case (Fig. 8a), a hotspot is observed close to the *Clock Tree* block which has the highest average heat generation rate when compared to the rest of the blocks. Moreover, the heat can diffuse easily to the other parts of the circuit due to the high thermal conductivity of the bulk Silicon. In addition to this, the peak temperature of the entire adder is not significantly greater than its colder parts. On the other hand, the resulting temperature map of FDSOI (Fig. 8b) shows that the generated heat is locally translated into temperature since the thermal conductance of the FDSOI structure is quite low at the heat generation spots. This is mainly due to the SiO_2 isolation layer, Si-SiO_2 thermal boundary resistance [15] and the reduced thermal conductance of silicon and

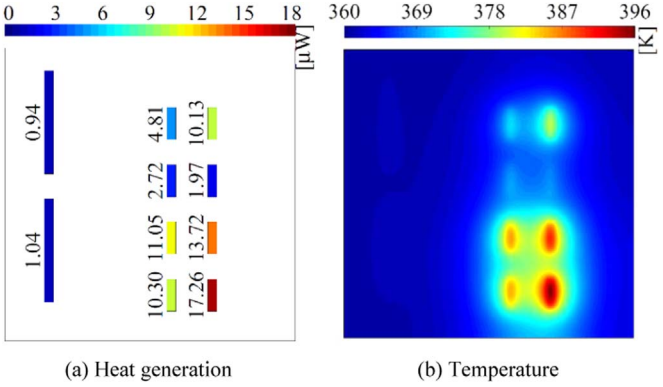


Fig. 9. Heat generation input of a high resolution thermal simulation and the resulting temperature map focused on the 48.6 μm≤x≤49.88 μm and 21.1 μm≤y≤22.38 μm portion of the 64-bit adder where the heat generation is very high. The spatial resolution used for thermal simulation is 1 nm^2 (a) Heat generation (b) Temperature.

SiO_2 thin films [25,26], which were previously explained. The thermal simulations for the bulk and the FDSOI geometries are repeated by using the block level heat generation inputs of Fig. 7. It is observed that the temperature distribution for the bulk case is almost exactly the same as the one of Fig. 8a. However, for the FDSOI case the temperature profile is quite different when compared to the one of Fig. 8b when the same experiment is performed with the heat generation input of Fig. 7. This result shows that the thermal simulations using block level heat generation inputs would give reliable results in bulk, however they would fail to detect the nanometer scale hotspots in FDSOI.

For observing the temperature profile of individual devices and the hottest spot in the entire block for the FDSOI case, it is needed to further increase the resolution of the simulation. For that, the location of the pixel with the highest temperature is determined on Fig. 6 and thermal simulations are repeated with a grid size of 1 nm^2 which is equal to the minimum layout pitch of the used technology. Fig. 9a shows the devices which are located on the spot with the highest temperature and their power dissipation values. Since the grid size (1 nm) is much smaller than the minimum gate length (40 nm), a significant area with zero heat generation can be observed (white area). The resulting temperature profile (Fig. 9b) indicates that the generated heat does not diffuse easily and each device heats up itself significantly. The temperature difference inside this small window (1.6 μm^2) is larger than 35 K and a temperature gradient of more than 100 K/μm can be observed. Moreover, without being too obvious, the same picture implies that a device with very high power dissipation might have a considerable impact on the temperature of its neighbouring device (observable from the four high power devices located on the lower right part of Fig. 9b).

To understand the effect of neighbour devices on each other, a heat generation scenario is created. Fig. 11a shows the locations of the devices with their names indicated on the left. The colours of the devices are determined according to the colour-bar on the top, which represents the average power density. The instantaneous and average power density of each device is defined by the following equations

$$PD_n(t) = \frac{P_n(t)}{W_n \bullet L_n} \tag{11}$$

$$\overline{PD_n} = \frac{1}{t_s}\int_0^{t_s} PD_n(t)dt \tag{12}$$

where W_n and L_n are the width and length of the corresponding device. For the test case of Fig. 11a, the sizes of the devices are selected according to the design rules of the utilized technology and the power density values are chosen according to the power density distribution of the adder circuit which is shown on Fig. 10. The power density of

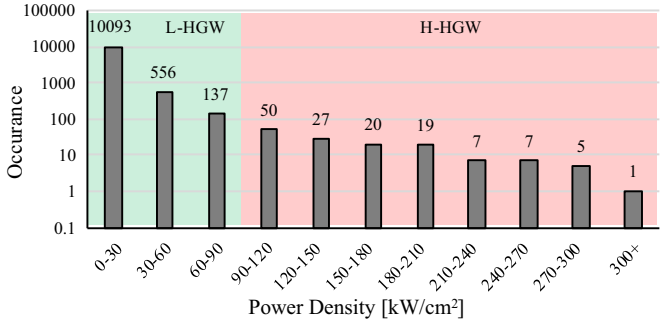


Fig. 10. Power density distribution of the transistors which belong to the 64-bit adder circuit.

Table 2
Parameters of the devices on Fig. 11.

Device	M1	M2-M11	M12
Power density [kW/cm^2]	42	300	0
Power dissipation [μW]	16.8	16.8	0
Width [μm]	1.00	0.14	0.14
Length [μm]	0.04	0.04	0.04

devices M2-M11 is roughly equal to the power density of the device which has the highest power density in the entire adder circuit according to the power dissipation analysis. Naturally, one of these devices should be responsible of the peak temperature of the overall implementation. On the other hand, M12 has the same size as M2-M11 with no power dissipation. Finally, the power dissipation of M1 is equal to the power dissipation of M2-M11 while its power density is much lower than the others due to its larger size. The horizontal distance between the devices is set to the minimum gate-to-gate distance permitted by the used design kit. All of the other parameters of each device are summarized on Table 2.

The resulting temperature profile of the listed heat generators is shown on Fig. 11b. The white dashed lines on Fig. 11b are the cut-lines which pass on the centre of the devices. The temperature waveforms occurring on these cut-lines are shown on Fig. 12. The peak temperature on cut-line 1 shows that the device with the highest power density (M2) can increase the temperature itself alone by 30 K. The peak temperature on cut-line 2, which is the peak temperature of the entire area, is observed on M7 with a temperature increase of more than 48 K when compared to the coolest point. Additionally, the temperature of M7 is 18 K higher than the temperature of M2 although they have the same heat density. This proves that the influence of the neighbouring devices (M3-M6 and M8-M11 in the example) can increase the peak temperature of the overall circuit significantly. Another observation which can be made on cut-line 2 is although M12 does not dissipate

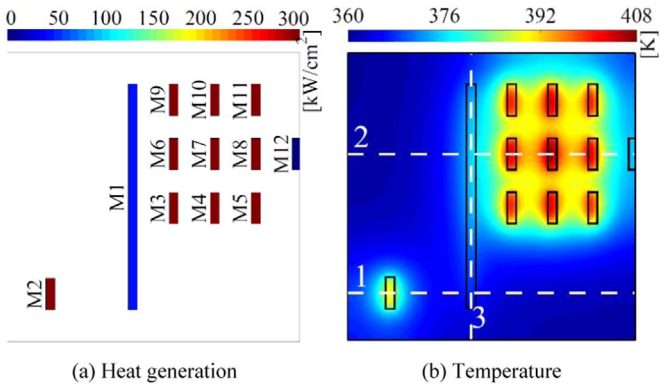


Fig. 11. Heat generation and the resulting temperature profile in a square with an area of 1.28×1.28 μm^2 (a) Heat generation (b) Temperature.

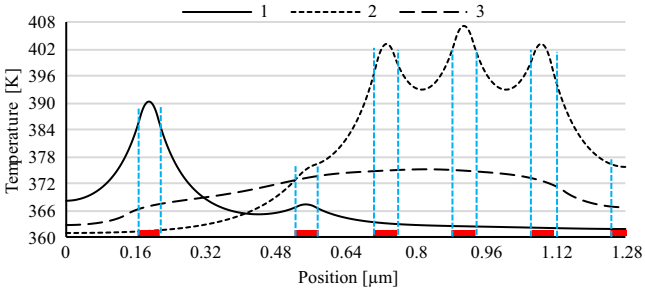


Fig. 12. Temperature profiles along the cut-lines (Fig. 11b) taken at the centre of the devices M2, M7 and M1. The red thick lines show the location of the devices on Fig. 11. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

any power, its temperature is 15 K higher than the lowest temperature in the area. This is primarily due to the self-heating of M3-M11 which are located very close to M12. The temperature of M1 is 15 K lower than the temperature of M2 although their power dissipation values are equal. This result proves that is the most substantial factor in terms of creating a nanometer scale hotspot in FDSOI is the power density rather than the power dissipation alone. Finally, the temperature of M1 shows more than 6.8 K variation along its width (cut-line 3 on Fig. 12). The gradient is due to the neighbouring devices (M3, M6, M9) which are very strong heaters. As a result of this variation, the threshold voltage and mobility of M1 would show different behaviours along its width. While this effects only the speed in digital circuits, it might create important mismatch problems in analog circuits.

The analysis shows that the hottest spot in FDSOI is most probably created by the self-heating of individual devices with the highest heat density. Moreover, their cumulative effect can increase the peak temperature significantly. Returning back to the 64-bit adder circuit, it can be seen that the power density range is limited by [0 kW/cm^2, 310 kW/cm^2] window (Fig. 10). However, the histogram shows an exponential decaying characteristic and most of the devices are cumulated in the region with very low heat generation density. At this point, we can define two windows on the histogram of Fig. 10 and call them *Low Heat Generation Window* (L-HGW) and *High Heat Generation Window* (H-HGW). The border between each other is quite arbitrary and defined as 90 kW/cm^2 for this case. What is important is while there are 10,786 devices in L-HGW, there are only 136 devices in H-HGW. This means that almost 98.8% of all the devices has a power density value less than 90 kW/cm^2 and only approximately 1.2% of the devices are generating more than 30% of the maximum power density observed in the entire circuit. This situation can also be observed on Fig. 6 where there are only very few red or yellow squares (due to the devices in H-HGW) and almost the entire adder area is covered with dark blue (due to the devices in L-HGW). Therefore, if the devices in this 1.2% portion are detected and modified in a way, the peak temperature of the block can be significantly decreased. The modification of the most critical devices with high power density can be performed by increasing their width with a trade-off of a slightly increased area and parasitic capacitances. Fig. 13 shows the histograms of the three cases where the width of the devices in H-HGW are kept constant (×1), doubled (×2) and tripled (×3). It can be seen that the power density distribution can be squeezed into a smaller window only by increasing the width of the devices which have a power density value of more than 90 kW/cm^2. The new maximum power density values are less than 170 kW/cm^2 and 125 kW/cm^2 for the ×2 and ×3 cases respectively. Although the maximum power density can be decreased significantly by performing the described analysis and modifications, it is quite cumbersome to perform an extensive power dissipation analysis by considering each device. Moreover, to have accurate results from the described analysis, one should perform the necessary simulations by considering the interconnect parasitic capacitances.

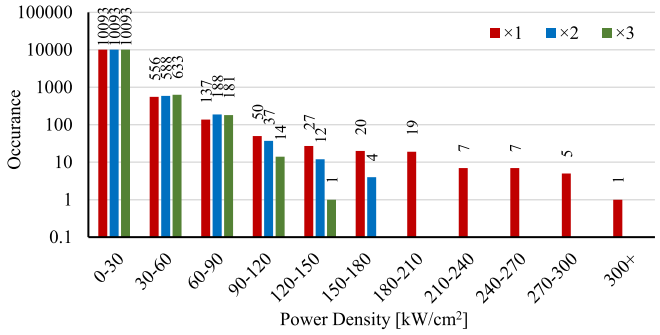


Fig. 13. The power density distribution histograms after increasing the width of the high power density devices by a factor of 2 and 3.

Consequently, the modifications on the critical devices in H-HGW have to be performed on top of a layout-ready design, which would increase the design time significantly. For that reason, a correlation between the devices in H-HGW and their functional properties is needed to be sought, so that the common properties of these devices and the reason of their high power density can be understood. This way, the necessary thermal design precautions can be taken during the early design stages rather than performing a very long analysis on an already finalized design.

4. Self-heating of devices with different functions

To observe the group of devices with different average power density values, the devices are categorized according to their functions. Considering the different functions in the circuit, there are 8 different groups. Each group is explained by means of Fig. 15, which shows the schematic of a simple logic function implemented in dynamic logic.

1) **Static (ST)**: The transistors of the static logic gates into which no clock signal enters (ST_1 and ST_2 on Fig. 15).
2) **Logic Tree (LT)**: The transistors used for implementing the Boolean functions of the dynamic logic gates (LT_1, LT_2 and LT_3 on Fig. 15).
3) **Footing Device (FD)**: The footing device which is used for preventing any leakage via the logic tree path during the pre-charge phase (FD on Fig. 15).
4) **Pre-charge Output Node (PO)**: The load device which is used to charge or discharge the output node during the pre-charge phase (PO on Fig. 15).
5) **Pre-charge Intermediate Node (PI)**: The devices used for pulling up or down the intermediate nodes in the logic tree of dynamic gates for preventing charge sharing problem (PI on Fig. 15).
6) **Clock Tree (CT)**: The devices of the static inverters which distribute the clock signal only to the dynamic logic gates (CT_1 and CT_2 on Fig. 15).
7) **True Single Phase Clock D Flip Flop (TSPC-DFF) (TS)**: The constitutive transistors of the TSPC-DFFs (Not shown on Fig. 15).
8) **C^2MOS D Flip Flop (C2)**: The constitutive transistors of the C^2MOS-DFFs (Not shown on Fig. 15).

The power density distribution of the entire 64-bit adder circuit (Fig. 10) is elaborated on Fig. 14 by providing individual distribution of each device group. It can be seen that the power density distribution of individual groups shows significant differences in terms of their mean values and population range. In the lowest power density range [0 kW/cm^2, 30 kW/cm^2], it is possible to find any group of devices and a huge percentage of the devices are populated in this window. However, in the high power density range [60 kW/cm^2, 330 kW/cm^2] which covers even a wider window than H-HGW, only 2 device groups out of 8 exists. These are the PO and PI groups. According to Table 3, the power density value of any of the devices other than the PO and PI groups is less than 55.2 kW/cm^2; while the devices in the PO and PI groups can reach to power density values more than 300 kW/cm^2. Moreover, the mean power density of the PO and PI devices is quite high when compared to the other groups. These results indicate that PO and PI are the most critical groups in terms of creating a hotspot in a circuit which is implemented in FDSOI. Consequently, instead of performing an extensive power dissipation analysis, one can focus on the sizing of these devices in the early stage of a design so that the local peak temperature values in a circuit can be decreased significantly (Fig. 15).

The high power density of the devices in the PO and PI groups can be understood by comparing the time domain behaviours of these devices with the devices from the other groups. Fig. 16 shows the time domain power density waveforms of the devices from the PO, PI, CT and LT groups. The evaluation (from t_1 to t_2) and the pre-charge (from t_2 to t_3) phases can be seen on Fig. 16. The devices which are located on any path between the input and the output are mainly active in the evaluation phase. Some of the examples of these devices are LT, ST and FD. Since these devices are on the data path, for a design with a short critical path delay, they should be sized in a way that they are ON (i.e. conducting) for a short period of time when compared to the entire evaluation period. Consequently, these devices are conducting current, hence generating heat for a short time. This can be seen on Fig. 16 by observing the power density waveform of the LT device which is ON for a very small time interval when compared to the other devices. The PO and PI devices are operating only during the pre-charge phase. However, they do not have to charge or discharge their corresponding nodes as fast as the devices on the data path. As long as they pre-charge the capacitance seen at their drain terminal before the beginning of the next evaluation phase, the circuit can safely work. Therefore, they are sized as small as possible for having a smaller area and less parasitic capacitance. The consequence is having a longer conduction time. Since the PO and PI devices are ON for longer durations when compared to the other groups, they have higher average power density values according to (12) and they become the most critical devices in terms of creating a hotspot.

The activity rate is an important parameter of power density in addition to conduction duration of a device. The CT devices have the highest activity rate among all the devices. The probability of conduction of a CT device is one in every clock period. However, for all the rest of the devices this probability is less than one and it depends on the input data, logic function and the location of the device in the circuit architecture. Nevertheless, the clock tree devices have less power density than the PO and PI devices. This is due to our previous reasoning that they have to provide a fast transition since they affect the critical path delay, hence the speed of the circuit. A fast transition brings less average power density (Fig. 16). Finally, it can be seen from Table 3 that the maximum and the mean power density values of PI is almost double the ones of PO although they are both conducting during the pre-charge phase and they have the same function. This is mainly due to the probability of discharging or charging an intermediate node during the evaluation phase that is higher than the probability of making an evaluation at the output node in most of the cases. This probability increases as the number of series devices decrease and the parallel branches increase between the corresponding node and the power supply in a logic tree (Fig. 15). Consequently, the temperature of the pre-charge devices become higher.

As it has been already mentioned, for having better thermal robustness, the maximum local temperature values should be decreased by decreasing the power density values of the critical devices (PO and PI). This can be done by increasing the width of these devices with a trade-off of an increased area and a slightly decreased speed similar to what has been done in Section. 3. Fig. 17 shows that the power density distribution can be squeezed into a smaller window only by increasing the width of the PO and PI devices by a factor of 2 and 3. It can be seen that the histograms of Figs. 13 and 17 are exactly the

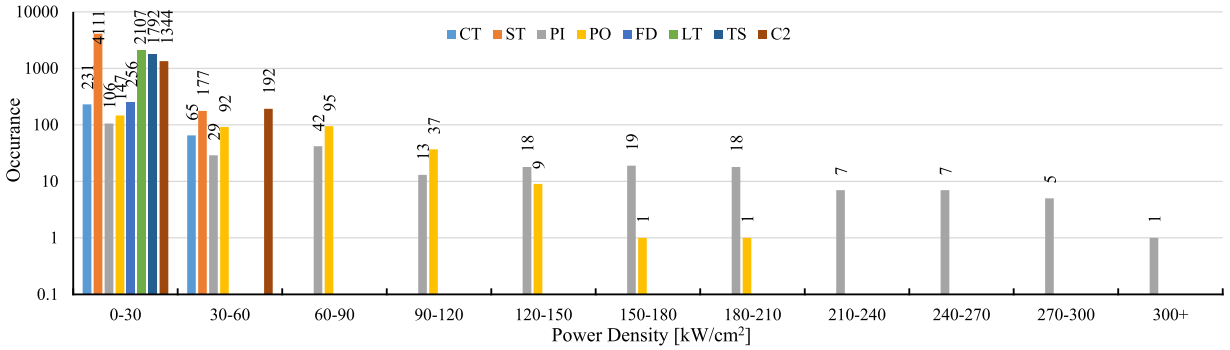


Fig. 14. Distribution of power density for different group of devices in the 64-bit adder circuit.

Table 3
Maximum and mean power density values and the number of devices of different device groups.

Group	Max [kW/cm^2]	Mean [kW/cm^2]	Number
PI	308.2	81.9	265
PO	181.0	47.4	382
ST	55.2	8.8	4288
CT	53.3	25.8	296
C2	52.5	10.9	1536
LT	26.3	3.9	2107
TS	15.4	4.8	1792
FD	12.2	6.2	256

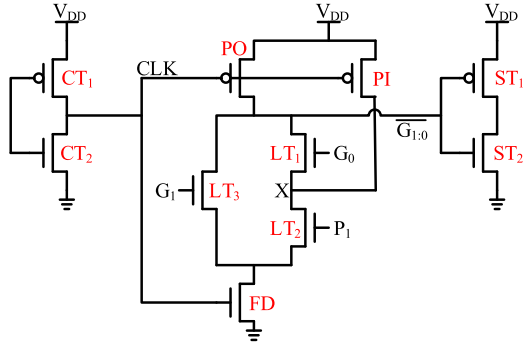


Fig. 15. The schematic which illustrates the devices belonging to different groups. The logic gate in the centre corresponds to the domino logic implementation of Generate-Merge function with N-type logic tree.

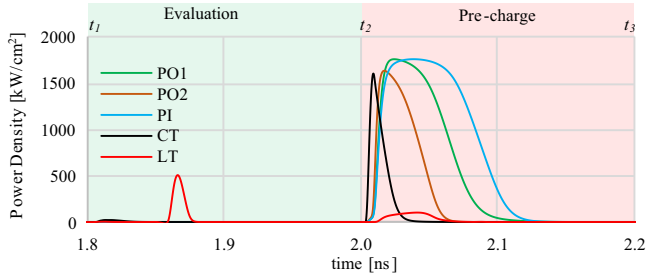


Fig. 16. Time domain power dissipation density waveforms of devices from different groups.

same when only H-HGW is considered and the maximum heat generation values are significantly decreased. Moreover, by knowing which group of devices are critical, the proper sizing can be done much more easily and the maximum heat generation value can be further decreased only considering a few devices in the entire circuit. In addition to proper sizing, the devices can be separated from each other to relax the heat flow [17]. Finally, thermal vias can be added to the drain ends of these devices similar to what is shown in [28] and [29] to

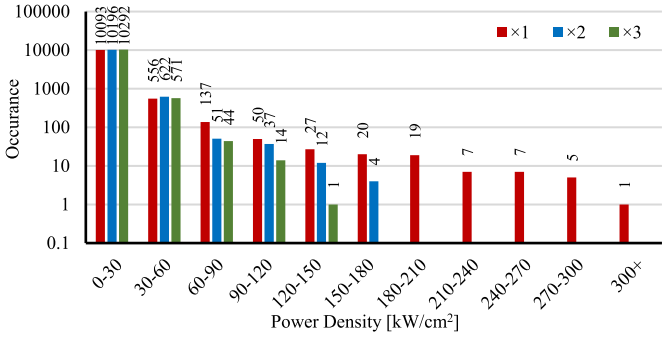


Fig. 17. The power density distribution histograms after increasing the width of the PI and PO devices by a factor of 2 and 3.

provide better heat diffusion paths, which will in turn increase the delay time.

5. Conclusion

In this paper, the thermal behaviours of very high speed digital blocks in bulk CMOS and FDSOI are studied and compared by experimenting a 64-bit parallel prefix adder operating under 900 mV power supply voltage. Thermal simulations on FDSOI show that the devices with high power density values are the main factors of the hotspots in a circuit due to the poor thermal conductance of FDSOI geometry. A single device with practical dimensions and power dissipation values can have a temperature that is 30 K higher compared to its surroundings only due to its self-heating. Moreover, closely placed devices can increase the temperature of their neighbours by 20 K. It is demonstrated that sizing the critical devices slightly larger 2–3 times) would decrease the maximum power density by 60% with a trade-off of a slight decrease in the speed and an increase in the area. Finally, it is shown that the devices which are responsible of pre-charging the intermediate and output nodes have the highest mean and maximum heat generation density and the same result can be obtained only by properly sizing these group of devices.

References

[1] D. Vasileska, K. Raleva, S.M. Goodnick, Heating effects in nanoscale devices, in: D. Vasileska (Ed.)Cutting Edge Nanotechnology, In-Tech, Janeza Trdine 9,51000 Rijeka, Croatia, 2010.
[2] M. Pedram, S. Nazarian, Thermal modeling, analysis, and management in VLSI circuits: principles and methods, Proc IEEE 94 (2006) 1487–1501.
[3] Y. Tsividis, C. McAndrew, Operation and Modeling of the MOS Transistor, 3rd ed., Oxford University Press, New York, 2011.
[4] S.S. Bhattacharya, S.K. Banerjee, B.Y. Nguyen, P.J. Tobin, Temperature-Dependence of the Anomalous Leakage Current in Polysilicon-on-Insulator Mosfets, IEEE Trans. Electron Dev. 41 (1994) 221–227.
[5] H.H. Su, F. Liu, A. Devgan, E. Acar, S. Nassif, Full chip leakage estimation considering power supply and temperature variations, in: Islped'03: Proceedings of the 2003 International Symposium on Low Power Electronics and Design, 2003, pp. 78–83.

[6] M.L. Mui, K. Banerjee, A. Mehrotra, Power supply optimization in sub-130 nm leakage dominant technologies, in: Proceedings of the Isqed 2004: 5th International Symposium on Quality Electronic Design, 2004, pp. 409–414.
[7] J.R. Black, Electromigration – a brief survey and some recent results, IEEE Trans. Electron Dev. Ed16 (1969) (338- &).
[8] P.C. Li, T.K. Young, Electromigration: the time bomb in deep-submicron ICs, IEEE Spectr. 33 (1996) 75–78.
[9] E. Pop, K.E. Goodson, Thermal phenomena in nanoscale transistors, In; Proceedings of the Itherm 2004, Vol 1, 2004, pp. 1–7.
[10] E. Pop, S. Sinha, K.E. Goodson, Heat generation and transport in nanometer-scale transistors, Proc. IEEE 94 (2006) 1587–1601.
[11] D. Vasileska, K. Raleva, S.M. Goodnick, Self-heating effects in nanoscale FD SOI devices: the role of the substrate, boundary conditions at various interfaces, and the dielectric material type for the BOX, IEEE Trans. Electron Dev. 56 (2009) 3064–3071.
[12] C.J. Ni, Z. Aksamija, J.Y. Murthy, U. Ravaioli, Coupled electro-thermal simulation of MOSFETs, J. Comput Electron. 11 (2012) 93–105.
[13] K. Raleva, D. Vasileska, A. Hossain, S.K. Yoo, S.M. Goodnick, Study of self-heating effects in SOI and conventional MOSFETs with electro-thermal particle-based device simulator, J. Comput Electron. 11 (2012) 106–117.
[14] J. Chen, G. Zhang, B.W. Li, Thermal contact resistance across nanoscale silicon dioxide and silicon interface, J. Appl. Phys. 112 (2012).
[15] E. Lampin, Q.H. Nguyen, P.A. Francioso, F. Cleri, Thermal boundary resistance at silicon-silica interfaces by molecular dynamics simulations, Appl. Phys. Lett. 100 (2012).
[16] E. Pop, Self heating and scaling of thin body transistors, in: Department of Electrical Engineering, Stanford University, 2004.
[17] J. Warnock, B. Curran, J. Badar, G. Fredeman, D. Plass, Y. Chan, S. Carey, G. Salem, F. Schroeder, F. Malgioglio, G. Mayer, C. Berry, M. Wood, Y.H. Chan, M. Mayo, J. Isakson, C. Nagarajan, T. Werner, L. Sigal, R. Nigaglioni, M. Cichanowski, J. Zitz, M. Ziegler, T. Bronson, G. Strevig, D. Dreps, R. Puri, D. Malone, D. Wendel, P.K. Mak, M. Blake, 22 nm next-generation IBM system z microprocessor, in: Proceedings of the Isscc Dig Tech Pap I, 58, 2015, pp. 70–U90.
[18] W. Huang, K. Skadron, S. Gurumurthi, R.J. Ribando, M.R. Stan, Differentiating the roles of IR measurement and simulation for power and temperature-aware design, in: Proceedings of the Int Sym Perform Anal, 2009, pp. 1–10.
[19] P.M. Kogge, H.S. Stone, Parallel algorithm for efficient solution of a general class of recurrence equations, IEEE Trans. Comput. C-22 (1973) 786–793.
[20] F. Gurkaynak, Y. Leblebici, L. Chaouat, P.J. McGuinness, Higher radix Kogge-Stone parallel prefix adder architectures, in: Proceedings of the Iscas 2000: IEEE International Symposium on Circuits and Systems, Vol V, 2000, pp. 609–612.
[21] Y. Shimazaki, R. Zlatanovici, B. Nikolic, A shared-well dual-supply-voltage 64-bit ALU, IEEE J. Solid-State Circuits 39 (2004) 494–500.
[22] J.R. Yuan, C. Svensson, P. Larsson, New domino logic precharged by clock and data, Electron. Lett. 29 (1993) 2188–2189.
[23] R. Zlatanovici, B. Nikolic, Power – performance optimal 64-bit carry-lookahead adders, Esscirc 2003 in: Proceedings of the 29th European Solid-State Circuits Conference, 2003, pp. 321–324.
[24] T.M. Tritt, Thermal Conductivity Theory, Properties, and Applications, 1 ed., Springer, US, 2004.
[25] T. Yamane, N. Nagai, S. Katayama, M. Todoki, Measurement of thermal conductivity of silicon dioxide thin films using a 3 omega method, J. Appl. Phys. 91 (2002) 9772–9776.
[26] S.M. Lee, D.G. Cahill, Heat transport in thin dielectric films, J. Appl. Phys. 81 (1997) 2590–2595.
[27] S. Im, N. Srivastava, K. Banerjee, K.E. Goodson, Scaling analysis of multilevel interconnect temperatures for high-performance ICs, IEEE Trans. Electron Dev. 52 (2005) 2710–2719.
[28] T.Y. Chiang, K. Banerjee, K.C. Saraswat, Effect of via separation and low-k dielectric materials on the thermal characteristics of Cu interconnects, in: Proceedings of the International Electron Devices Meeting 2000, Technical Digest, 2000, pp. 261–264.
[29] Z. Wang, G. Dong, Y.T. Yang, J.W. Li, Effect of dummy vias on interconnect temperature variation, Chin. Sci. Bull. 56 (2011) 2286–2290.

Can Baltacı received the B.S. degree in Electronics Engineering from Middle East Technical University, Ankara, Turkey and M.S. degree in Electrical Engineering from Swiss Federal Institute of Technology in Lausanne (EPFL), Switzerland in 2010 and 2013, respectively. He is currently studying towards his Ph.D. degree at the Swiss Federal Institute of Technology in Lausanne (EPFL). His research interests include self-heating effects in high performance analog and digital circuits implemented in bulk and FDSOI.

Yusuf Leblebici (M90SM98F09) received his B.Sc. and M.Sc. degrees in electrical engineering from Istanbul Technical University, in 1984 and in 1986, respectively, and his Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign (UIUC) in 1990. Since 2002, Dr. Leblebici has been a Chair Professor at the Swiss Federal Institute of Technology in Lausanne (EPFL), and director of Microelectronic Systems Laboratory. His research interests include design of high-speed CMOS digital and mixed-signal integrated circuits, computer-aided design of VLSI systems, intelligent sensor interfaces, modelling and simulation of semiconductor devices, and VLSI reliability analysis. He is the coauthor of 6 textbooks, as well as more than 300 articles published in various journals and conferences. He has served as an Associate Editor of IEEE Transactions on Circuits and Systems (II), and IEEE Transactions on Very Large Scale Integrated (VLSI) Systems. He has also served as the general co-chair of the 2006 European Solid-State Circuits Conference, and the 2006 European Solid State Device Research Conference (ESSCIRC/ESSDERC). He is a Fellow of IEEE and has been elected as Distinguished Lecturer of the IEEE Circuits and Systems Society for 2010–2011.